Understanding Google Analytics (GA) Limitations and Caveats

Every Web analytics tool has its own unique capabilities, as well as its own limitations, and Google Analytics (GA) is no different. When using GA to collect and analyze data it is important to understand certain caveats in order to gain richer insights.

Historical Data

EPA's GA profile began collecting data in March 2013. However, GA retains historical data for up to three years, so you will be able to access months or years worth of data as it continues to populate in the agency profile. Nevertheless, the EPA Web Analytics Program recommends that you retain a copy of all reports you generate in GA.

Missing Page Tags

If the GA script is removed from a Web page, either because one of the core JavaScript files or the Google Tag Manager (GTM) script that contains the agency's GA script is removed, GA will no longer collect metrics for that Web page. Once the GA script is added back to the Web page, the metrics lost in the interim cannot be recovered. This is why the agency GA script is a requirement for all EPA public Web pages. If you believe any of your office's public Web pages are not tagged with the GA script, please contact Bronson.Samuel@epa.gov

Sampling

"Sampling in Google Analytics or in any web analytics software refers to the practice of selecting a subset of data from your website traffic. Sampling is widely used in statistical analysis because analyzing a subset of data gives similar results to analyzing all of the data. In addition, sampling speeds up processing for reports when the volume of data is so large as to slow down report queries." – Google: https://developers.google.com/analytics/resources/concepts/gaConceptsSampling Exit.

Standard GA reports rely on data tables that are precompiled and can therefore be processed incredibly fast without the need for sampling, in most cases. However, by using some of the more advanced features of GA, including custom segments and custom reports, you can trigger data sampling. This is because you are requesting Google to process data in a non-standard way, which requires a lot more processing.

When a report is being sampled, a yellow rectangular box will appear in the top right corner that shows the number and percentage of Visits on which the report is based. Samples can be quite high in many instances, and therefore extremely accurate. Try reducing the number of segments or shortening the time period of the request to increase the sample size.

Cookie Deletion and Return Visitors

A small percentage of Visitors to EPA will delete their Web browser's cookies prior to their next Visit. With the GA cookies deleted, these Visitors will be counted as New Visits upon their return to the EPA website. This is the known and accepted reality among Web analytics. Therefore, when you interpret the Return Visits metric in GA, consider the metric to represent the Minimum Return Visits.

Some Return Visits will inevitably be lost due to cookie deletion, but as is true of most Web traffic metrics, it is the trend over time that will provide the most insight. Log file analyzers do not offer a good alternative in this regard, as they rely on IP addresses to identify Return Visits, and large companies often have dynamic IP addresses that can change between Visits or even during a Visit.

How does GA Track Downloads, like PDF Documents?

GA, like all page tagging tools, does not track actual file downloads, but instead tracks the number of clicks on links to downloadable files. However, for EPA files to be tracked they must reside on the EPA website so the action can trigger the agency GA script. Clicks on links to EPA files that reside on external websites will not be tracked. This is the main distinction from how file downloads are calculated by log file analysis tools, which process all requests in the server logs. On the other hand, log files often have inflated pdf counts because pdfs can trigger multiple server requests as the user scrolls down the document.

EPA's custom implementation of GA automatically tracks external links, email links, and links to many common file types, including doc, docx, exe, pdf, ppt, pptx, mp3, and zip files. It is also possible, using Event Tracking, to implement custom tags on other types of links, images, and video files. However, it is critical that the Agency maintains consistent naming conventions for all Event Tags. If you are considering Event Tracking that goes beyond the auto-tracking included in the agency GA script, please contact Bronson.Samuel@epa.gov.

"Other" in Content Reports

Depending on the amount of content generated by a GA report, you may see "other" among the list of URLs.

GA has a daily limit to how many rows of data it can process in a report table: 50,000 rows for the free version and 3,000,000 rows for the premium version. Once that limit is reached, the data from the remaining rows are aggregated under the title "other." When this occurs in the standard content reports, for example, "other" represents the remaining URLs after the threshold for data rows is reached.

In general, the premium version of GA has enough processing power to generate complete reports, even at the agency level, but creating separate profiles for certain website content is another way to avoid data table limitations.

"Not provided" in Organic Search Traffic Report

The Organic Search Traffic Report provides the keywords users entered in search engines in order to reach the EPA website. However, Google does not provide the keywords of users who were signed into another Google product (e.g. Gmail, YouTube, etc.) when they executed their search on the Google

Search Engine. This is not a limitation particular to Google Analytics; Google encrypts the keyword data, so it is not accessible by any Web analytics tool. In GA, the encrypted Google keywords are represented under the "not provided" row in the Organic Search Traffic Report.

You can access the full list of organic keywords entered in the Google Search Engine with the Google Webmaster Report that has been integrated with the EPA agency GA account. The Google Webmaster data is located in the GA Search Engine Optimization Report. However, this report will always provide the full list of keywords used to reach *any* EPA website, regardless of whether you view the data from the main agency profile or a separate profile. In other words, you cannot get the full list of keywords used to reach just your Web area.

"Not set" in GA Reports

When GA has not received any information for a dimension, the term "not set" will appear in GA reports. This can occur for a variety of reasons, such as you requesting a custom report with metrics and dimensions that are incompatible. If you need assistance with a report, please contact Bronson.Samuel@epa.gov.

Self-Referrals

Some referral pages will be listed as EPA Web pages in EPA's GA account. The main reason for this is that not all EPA Web pages were updated with the agency's GA code at the same time. When Visitors view a Web page where the GA code is not embedded and then subsequently view a page that does include the GA code, the previous page is treated as a referral page.

Even after all EPA pages are updated with the agency GA code, Return Visitors who previously entered the site from a Web page that contained no GA code may still be attributed with referrals from EPA Web pages because of the GA cookie in their browser. The GA cookie maintains the original referral source for Return Visitors who return through direct methods (bookmarks or typing-in a URL). The Visitor would either have to delete the original cookies, wait for them to expire, or Visit the website through a search engine or non-EPA referral page. Eventually, one of these three scenarios will occur for all those who visited the website during our implementation stage. Therefore, the self-referral count will decrease over time.

Time on Page and Visit Duration Calculations

Time on Page represents the average amount of time, in seconds, a Visitor spends on a particular Web page. Visit Duration represents the average time, in seconds, for an entire Visit. However, like other Web analytics tools, it is difficult to capture the exact time for every Visit because of how these metrics are calculated. Technically, Time on Page represents the time between the start time of a specified Pageview and the start time of a subsequent Pageview or Event (**see the Glossary of Terms for metric definitions**). However, Visits that include only one Page View and no subsequent Events will have a Time on Page of zero seconds (for example, a Visitor views a page and then closes the browser). Time on

Page calculations of zero seconds do contribute to the average, so you should interpret Time on Page carefully.

Visit Duration technically represents the time between the start time of the first Pageview and the start time of the last Pageview of a Visit. Visits with only one Pageview may be attributed a Visit Duration of zero seconds, which contributes to the average. Additionally, the time spent on the last Web page will not be fully captured. Therefore, you should also interpet Visit Duration carefully. Consider measuring Visit Duration for Visits of two or more Pageviews to eliminate Visit Durations of zero seconds.

Visit-Based Metrics are only Accurate at the Profile-Level

A Visit represents the full timespan, or session, that a Visitor spends on a particular website, starting when the first Web page of that site is loaded in the browser and ending when the Visitor either leaves the website, closes their browser, or the Visit times out after 30 minutes of inactivity.

Visit-based metrics (e.g. Visits, Return Visits, Duration of Visit, etc.) are most useful for understanding Web traffic to an entire website, subdomain, or Web application; or else to a large collection of interconnected Web pages that represent a single topic or content type. For single pages, or groups of unrelated pages, users should focus on page-level metrics, such as Pageviews, Unique Pageviews, and Time on Page.

It is important to understand that accurate Visit-based metrics can only be generated at the profile-level. Therefore, profiles should be reserved for large sites and applications.

When you filter within a profile, down to the directory- or page-level, you should ignore all Visit-based metrics. For example, using custom reports, it is possible to generate a report that shows the number of Visits for a single directory or page. However, in this context, “Visits,” will actually mean Entrances, and metrics like Duration of Visit, New Visits, and other Visit-based metrics will not be accurate.

Once you filter content within a profile, in order to access metrics on a subset of pages, only page-level metrics will be accurate.

Tip: At the page-level, Unique Pageviews is akin to Visits and can be used in its stead.

To learn more about all the GA metrics, visit the **Glossary of Terms**.

Dynamic URLs with Query Parameters

Many EPA applications output dynamic URLs with appended query parameters. Sometimes, you may want metrics for these complete URLs in order to analyze the exact queries that are executed. However, because each unique query produces a unique URL, you may wish to remove the query parameters and allow GA to aggregate all executed queries in GA reports.

Upon request, we can remove query parameters from your dynamic URLs reported in GA so they can be automatically aggregated in the agency profile. Even after these parameters are removed in the agency profile, it is still possible to retrieve the full dynamic URLs and the actual URLs will remain the same for Visitors to the EPA website.

If you are interested in removing the query parameters from your dynamic URLs within GA reports, please contact Bronson.Samuel@epa.gov.

Pages Reported with /index.html

Many Web pages on www.epa.gov can generally be requested as a directory or a full index page:

1. …/directory/
2. …/directory/index.htm or …/directory/index.html (depending on the file extension)

This URL duplication is reflected in GA reports as well, which can make gathering metrics for individual pages tedious.

To correct for this duplication, the agency GA profile includes a fix that combines these two URL versions and represents the aggregated data for each page with the full index version (e.g. epa.gov/index.html). However, this fix also causes GA to report some URLs inaccurately, because the directory does not contain a file named index.htm(l). The metrics for these URLs are still accurate, though, and we also maintain a profile where this fix is not implemented and all URL versions are unchanged.

Aliases

Before accessing metrics in GA, you should know what aliases exist for your content because *GA will not automatically aggregate aliases*. For this reason, creating multiple aliases is generally considered a bad practice from a Web analytics perspective.

To correct for this problem, you can redirect all your aliases to the primary alias. In EPA's Drupal WebCMS, Web areas generally have no more than one alias and it redirects to the primary alias as a rule. For non-Drupal content, you should consider retiring URL aliases or redirecting them to the primary alias. To learn more about existing EPA aliases, visit Search Central: http://nlquery.epa.gov/.

Filtering out Search Results Pages

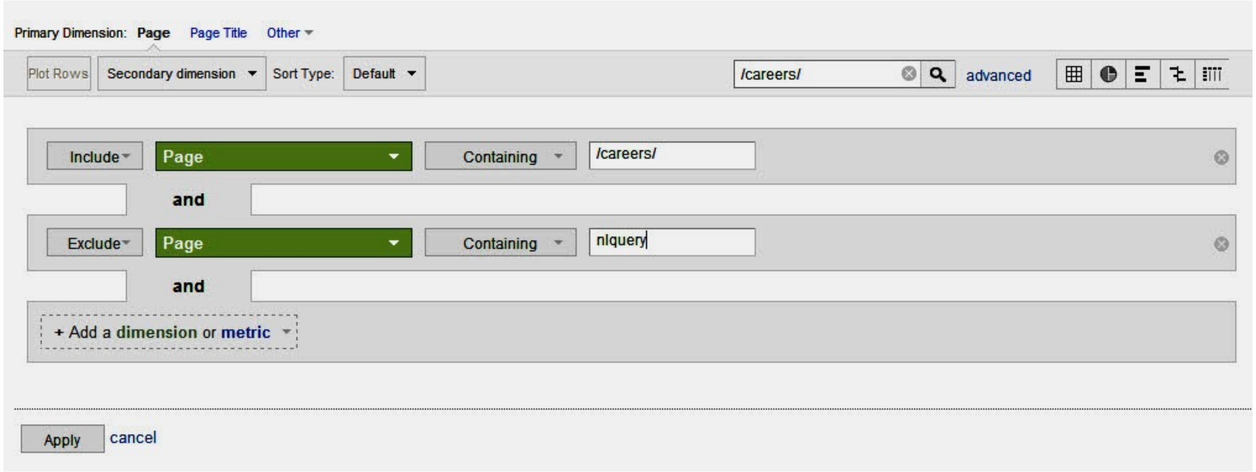
There are generally two ways to identify and organize Web pages in GA reports: by URL and by Page Title. The most common approach to creating reports is to filter by characters in URL strings, such as a root directory. For example, you might access the All Pages Report in the content section of the agency GA profile and then filter that report for the /careers/ directory. The results will include all URLs that include "/careers/," *including Search Results pages from searches executed on a /careers/ page*.

To filter out Search pages from a report, follow these steps:

1. Click the "advanced" button next to the search box located at the top of the report table
2. Select "Exclude" as the first action
3. Select "Page" as the dimension

4. Select “Containing” as the second action
5. Enter “nlquery” into the filter box
6. Click apply

See the screenshot below for reference:



/careers/

Include Page Containing /careers/

and

Exclude Page Containing nlquery

and

The same filtering process can be used to remove other pages that you might not want included in a report. For example, some root directories may also serve as subdirectories for URLs in other program offices. It is important to review every report to ensure that only the desired content is included.